Word Sense Disambiguation using Heterogeneous Language Resources

Kiyoaki Shirai **Takayuki Tamagaki**
School of Information Science
Japan Advanced Institute of Science and Technology
{kshirai,t-tamaga}@jaist.ac.jp

Abstract

This paper proposes a robust method for word sense disambiguation of Japanese. We combined several classifiers using heterogeneous language resources, a machine readable dictionary and a word sense tagged corpus. According to our experimental results, our method outperformed the best single classifier for recall and applicability.

1 Introduction

Word sense disambiguation (WSD) is the process of selecting the appropriate meaning or sense for a given word in a document. Obviously, WSD is one of the fundamental and important processes needed for many natural language processing (NLP) applications, such as machine translation systems. Over the past decade, many studies have been made on WSD of Japanese (Fujii et al., 1996; Shinnou, 2002; Shinnou and Sasaki, 2002). Most current research used machine learning techniques (Li and Takeuchi, 1997; Murata et al., 2001), and achieved good performance. However, as supervised learning methods require word sense tagged corpora, they often suffer from data sparseness, i.e., words which do not occur frequently in a training corpus can not be disambiguated. Therefore, we cannot use supervised learning algorithms alone in practical NLP applications, especially when it is necessary to disambiguate both high frequency and low frequency words.

This paper aims at developing a robust WSD system for Japanese words, and proposes a method which combines several classifiers for WSD, classifiers learned from a sense-tagged corpus and those obtained from a machine readable dictionary (MRD). The main purpose of combining several classifiers derived from heterogeneous language resources (word sense tagged corpus and MRD) is to increase the recall and applicability of the overall WSD system. Even when a classifier obtained by supervised learning can not determine the correct meaning for a certain word due to lack of training data, the classifiers from an MRD may be able to determine the correct sense. Thus, the robustness of the WSD system is improved by using several classifiers simultaneously.

2 Our Method

In this paper, word senses or meanings are defined according to the Japanese dictionary, the *Iwanami Kokugo Jiten* (Nishio et al., 1994).

The basic idea of our method is to combine the following four classifiers.

1. Classifier using example sentences in an MRD
2. Classifier using grammatical information in an MRD
3. SVM (Support Vector Machine) classifier
4. Baseline classifier

Notice that classifiers 1 and 2 use an MRD (the *Iwanami Kokugo Jiten*), while 3 and 4 use a sense-tagged corpus. Thus two kinds of language resources are used for WSD.

2.1 Classifier using Example Sentences in an MRD

2.1.1 Overview

In the Iwanami Kokugo Jiten, word definitions often contain example sentences. Figure 1 shows several such example sentences in the sense set of the Japanese verb "*aisuru*" (love). In Figure 1, the sentences in square brackets are examples, in which the headword is indicated by boldface.

aisuru

1) to have strong feelings of affection for someone/something
   [*ko o* ***aisuru*** (He/She loves his/her child)] (**E1**)
   [*kuni o* ***aisuru*** (He/She loves his/her nation)] (**E2**)
2) to have very strong feelings of affection for someone that you are sexually attracted to
3) to like or enjoy something very much
   [*sake o* ***aisuru*** (He/She loves drinking)] (**E3**)

Figure 1: Sense Set of "*aisuru*" (love)

shitau

1) to follow someone/something with full of affection
   [*haha o* ***shitau*** (He/She is attached to his/her mother)] (**E4**)
   [*kokoku o* ***shitau*** (He/She is attached to his/her home country)] (**E5**)
   [*kanojo ga hisoka ni* ***shitau*** *seinen* (Young man she loves in secret)] (**E6**)
2) to respect one's virtue, education or skills
   [*toku o* ***shitau*** (He/She respects one's virtue)] (**E7**)

Figure 2: Sense Set of "*shitau*" (be attached to)

The WSD classifier described here measures the similarity between an input sentence containing a target word and example sentences in an MRD, selects the example sentence which has the highest similarity and outputs the word sense that contains it in its definition. For example, let us consider the case where the word sense of the verb "*aisuru*" in the sentence **S1** should be disambiguated.

S1 *kare* (he) *wa* (TOP) *musume* (daughter) *o* (ACC) *aisuru* (love)
(He loves his daughter.)

Notice that cases are indicated by case-markers in Japanese such as "*o*" (accusative case-marker) and "*wa*" (topical case-marker). The classifier measures the similarity between this sentence and the example sentences **E1**, **E2** and **E3** in Figure 1 [1]. Among them, **E1** may have the highest similarity with **S1**. Therefore, the classifier selects sense 1) in Figure 1 as the correct meaning.

2.1.2 Extraction of Example Sentences from Sense Descriptions of Hypernyms

One of the problems in using example sentences from the Iwanami Kokugo Jiten is that the number of example sentences in the dictionary is not large enough. This may cause a data sparseness problem, especially since not all definitions in the dictionary contain example sentences. For instance, there is no example sentence for definition 2) in Figure 1. Such meanings will never be selected by the classifier. To overcome this problem, example sentences in the definitions of hypernyms are also used for WSD. We assume that the hypernym of a verb is the last verb of a definition sentence. For example, the following is the original definition of sense 2) of "*aisuru*" in Japanese.

aisuru

2) *isei* (opposite sex) *o* (ACC) *koi* (love) *shitau* (be attached to)

In this case, the last verb, "*shitau*" (be attached to) is assumed to be the hypernym of the verb "*aisuru*" (love). Therefore, the example sentences **E4**, **E5**, **E6** and **E7**, which are in the definition of "*shitau*" as shown in Figure 2, are extracted as example sentences for sense 2) of "*aisuru*". In this way, we can obtain example sentences for those senses for which no example sentence is given.

2.1.3 Sentence Similarity

In this paper, instead of the similarity between an input sentence and an individual example sentence, the similarity between an input sentence s and a set of example sentences E for each sense, $sim(s, E)$, is considered.

$Sim(s, E)$ is defined according to the similarity between two case-filler nouns of the same case. First, NE_c and $NE_{c'}$ are extracted for each sense from an MRD. NE_c is the set of case-filler nouns extracted from example sentences, where c is a case-marker such as *o* (ACC) and *ga* (NOM). $NE_{c'}$ is the set of case-fillers extracted

[1] Subjects (he/she) are omitted in these sentences.

1) NE_o = { *ko* (child), *kuni* (nation) }
2) $NE_{o'}$ = { *haha* (mother), *kokoku* (home country), *toku* (virtue) }
   $NE_{ga'}$ = { *kanojo* (she) }
3) NE_o = { *sake* (alcohol) }

Figure 3: Extracted Case-fillers for "*aisuru*"

from example sentences in the definition of hypernyms. For example, for sense 1) of "*aisuru*" in Figure 1, *ko* (child) and *kuni* (nation) are accusative case-fillers of the verb *aisuru* (love) in **E1** and **E2**, respectively. Thus NE_o for the sense 1) is $\{ko, kuni\}$. For sense 2) of "*aisuru*", *haha* (mother), *kokoku* (home country) and *toku* (virtue) are accusative case-fillers in **E4**, **E5** and **E7**, respectively. As **E4**, **E5** and **E7** are example sentences of the hypernym of sense 2), these nouns are members of $NE_{o'}$ for sense 2). The case-fillers for the other cases are extracted in the same way. The extracted case-fillers for all senses of the verb *aisuru* (love) are summarized in Figure 3.

Next, $Sim(s, E)$ is defined as the equation (1)

$$Sim(s,e) = \sum_c w_c \cdot s_c(ns_c, NE_c) \quad (1)$$

$$s_c(ns_c, NE_c) = \max_{ne_c \in NE_c} s(ns_c, ne_c) \quad (2)$$

$$s(w_i, w_j) = \frac{2 \times d_k}{d_i + d_j} \quad (3)$$

In (1), $s_c(ns_c, NE_c)$ is the similarity between a case-filler ns_c of a case c in a sentence s and a set of case-fillers NE_c of the same case c extracted from example sentences in an MRD, which is given by equation (2). w_c in (1) is a weight parameter for the case c, which is defined empirically. We set weights $w_{c'}$ to be smaller than w_c, where case c' means that case-fillers are extracted from the example sentences of a hypernym of a verb, while c refers to the example sentences of the verb itself. In (2), $s_c(ns_c, ne_c)$ is the similarity between two nouns, ns_c and ne_c. It is defined by a thesaurus as equation (3). In (3), d_i and d_j are the depth of words w_i and w_j in a thesaurus, respectively, and d_k is the depth of the common superior class of w_i and w_j. For this study, we used the Japanese thesaurus *Nihongo Goi Taikei* (Ikehara et al., 1997) to calculate $s(w_i, w_j)$.

sarani
1) still more, further, furthermore
2) ⟨⟨with a negative expression⟩⟩ not in the least, not at all

Figure 4: Sense Set of "*sarani*" (more)

2.2 Classifier using Grammatical Information in an MRD

The second classifier uses grammatical information in an MRD. In the Iwanami Kokugo Jiten, grammatical constraints for a certain word sense are sometimes described. For example, see Figure 4, the sense set of the Japanese adverb "*sarani*" (more) in the Iwanami Kokugo Jiten. The description in double brackets ("⟨⟨" and "⟩⟩") is the grammatical information for sense 2), i.e., the adverb *sarani* whose meaning is sense 2) always appears with a negative expression.

Let us consider the sentence **S2**.

S2 *kôkai* (regret) *nado* *sarani* *si nai* (not)
(He/She doesn't regret it at all)

We can guess the correct sense of the adverb "*sarani*" in **S2** is sense 2) in Figure 4, because there is the negative expression *nai* (not). In this way, grammatical information in an MRD can provide effective clues for WSD.

We developed the WSD classifier using grammatical information. First, we regard grammatical information as conditions that an input sentence should satisfy. The classifier checks whether an input sentence satisfies the conditions described by grammatical information for all meanings, and outputs all of those meanings which pass the check. Otherwise, the classifier outputs nothing, i.e., it can not determine the correct meaning.

As described earlier, grammatical information is described in double brackets in the Iwanami Kokugo Jiten. We extracted such descriptions, and developed a system which judges whether or not a sentence satisfies the conditions defined by the grammatical information. Followings are types of such conditions.

- Condition of inflection
- Condition of a headword, POS(part-of-speech) or conjugation form of the word just

before or after the target word

- Condition of an idiom
- Condition of a negative expression
- Condition of a position of a word in a sentence

There are 973 definitions containing grammatical information in the Iwanami Kokugo Jiten. Out of the 973, our classifier can handle grammatical information for 582 senses. As the number of meanings of polysemous words in our dictionary is 37,908, the classifier using grammatical information can handle only 1.5% of them. The applicability of this classifier thus appears to be quite low. However, since many common words include grammatical information in the dictionary, we believe that the classifier is actually more applicable than expected. Furthermore, grammatical information is a reliable feature for WSD, and it appears that the correct word sense is mostly selected when this classifier is applied. For such reasons, when this classifier is combined with other classifiers, it makes a positive contribution to the performance of the overall WSD system.

2.3 SVM Classifier

The third classifier is the SVM classifier, one of the classifiers based on supervised learning. The features used in the model include POSs / surface forms of words just before and after the target word, base forms of content words found in $\pm n$ word window [2], and so on. We used the LIBSVM package [3] for training the SVM classifier. The SVM model is $\nu-$SVM (Schölkopf, 2000) with the linear kernel, where the parameter $\nu = 0.0001$. The pairwise method is used to apply SVM to multi classification.

The RWC corpus (Hasida et al., 1998) is used as the training data. It is made up of 3,000 newspaper articles extracted from the 1994 Mainichi Shimbun, consisting of 888,000 words. Out of 3,000 newspaper articles, we use 2,400 articles for training. SVM classifiers are trained for 2,084 words which occur more than 10 times in the training data. No meaning is selected by the SVM classifier for the other words.

[2] Here we set n to 20.

[3] `http://www.csie.ntu.edu.tw/%7Ecjlin/libsvm/`

Table 1: Correctness of each classifier on the validation data

	EXAM	GRAM	SVM	BL
C_{all}	0.329	0.816	0.797	0.796

2.4 Baseline Classifier

The last classifier is the baseline classifier which always selects the most frequently used meaning. When there is more than one meaning with equally high frequency, the classifier chooses one of the meanings randomly. This is the typical baseline model when using only the word sense tagged corpus for WSD.

2.5 Combined Model

The combined model is the WSD system using the four classifiers described in 2.1, 2.2, 2.3 and 2.4. In this subsection, we will describe how to combine these classifiers.

First, we prepare validation data, a sense-tagged corpus, as common test data for the four classifiers. The performance of the classifiers for a word w is evaluated by *correctness* C_w defined by (4).

$$C_w = \frac{\text{\# of words in which one of meanings selected by a classifier is correct}}{\text{\# of words for which a classifier selects one or more meanings}} \quad (4)$$

As mentioned earlier, the main reason for combining several classifiers is to improve the recall and applicability of the WSD system. Note that a classifier which often outputs a correct meaning would achieve high correctness C_w, even though it also outputs wrong meanings. Thus, the higher the C_w of a classifier, the more it improves the recall of the combined model.

Combining the four classifiers is a simple process. The correctness, C_w, of each classifier for each word w is measured on the validation data. When more than two classifiers output meanings for a given word, their C_w scores are compared. Then, the word senses provided by the best classifier are selected as the final outputs.

When the number of words in the validation data is small, comparison of the classifiers' C_w is unreliable. For that reason, when the number of words in the validation data is less that a certain threshold O_h, the correctness for all words

in validation data (C_{all}) is compared, rather than comparing the correctness for individual words w (C_w). In the experiment in section 3, we set O_h to 10.

3 Evaluation

In this section, we will describe the experiment to evaluate our proposed method. Out of 3,000 newspaper articles in the RWC corpus, 300 articles was used as validation data, and other 300 articles as test data. They were mutually exclusive with the training data used for training the SVM and baseline classifier. Only polysemous words in the corpus were disambiguated. The number of such target instances in the validation and test data was 13,819 and 13,494, respectively.

Table 1 shows the correctness of each classifier for all words in validation data. "EXAM", "GRAM", "SVM" and "BL" represents the classifier using example sentences, the classifier using grammatical information, the SVM classifier, and the baseline classifier, respectively. The best classifiers according to C_{all} on the validation data is "GRAM".

Table 2 reveals the precision, recall, F-measure [4] and applicability of the combined model and the single classifiers (EXAM, GRAM, SVM and BL) on the test data. "Applicability" indicates the ratio of the number of instances disambiguated by a classifier to the total number of target instances.

Previous papers (Takamura et al., 2001; Murata et al., 2001) have reported that the SVM classifier performed well, but in our experiment its precision was almost same as that of the baseline classifier. We do not understand the precise reason for that, but will examine the effective features used for the SVM classifier to improve it in future.

The combined model outperformed any single classifier for recall and applicability. This indicates that our goal — to improve the recall and applicability of the WSD system by combining several classifiers — was accomplished to some degree. On the other hand, the precision and F-measure of the combined model was less than that of the SVM and baseline classifier, which were the best among single classifiers. This was because the precision of the classifiers using example sentences (EXAM) and grammatical information (GRAM) was low. To improve the precision of these classifiers using an MRD is an important future project.

In the combined model, 71.0% of target instances were disambiguated by the classifier trained on the sense-tagged corpus (52.6% by SVM and 18.4% by BL), while 23.4% were disambiguated by the classifier using an MRD (1.62% by EXAM and 21.8% by GRAM). This indicates that both a sense-tagged corpus and an MRD, i.e., the RWC corpus and the Iwanami Kokugo Jiten in this experiment, were useful for WSD.

4 Related Works

This paper proposes a method for combining several classifiers for Japanese WSD. In this section, some previous research using ensemble of two or more WSD classifiers will be compared with our proposed method.

Several research (Pedersen, 2001; Takamura et al., 2001; Murata et al., 2001; Klein et al., 2002) proposed combining several classifiers trained on the same training corpora with different feature sets. One of the characteristics of these methods was that only a word sense tagged corpus was used as a knowledge resource. Therefore, the ensemble of several classifiers appeared to improve the precision of the overall WSD system, but not its recall and applicability.

Methods that combined classifiers using sense-tagged corpora and other language resources were also proposed. For example, Agirre et al. proposed combining classifiers using machine learning techniques and classifiers based on the WordNet (Agirre et al., 2000). Instead of a thesaurus, this study used an MRD as a language resource in addition to a sense-tagged corpus.

Litkowski proposed the method combining classifiers trained on a sense-tagged corpus and an MRD (Litkowski, 2002). However, his combination of two classifiers was indirect. Since the word sense definitions of the two classifiers were different, he converted the word senses produced by the MRD classifier to those defined by the classifier using the sense-tagged corpus. Such conversion is not always successful. Our approach, on the other hands, requires no sense conversion:

[4] $\frac{2PR}{P+R}$ where P and R represents the precision and recall.

Table 2: Results

	Precision	Recall	F-measure	Applicability
Combined	0.724	**0.772**	0.747	**0.944**
EXAM	0.466	0.080	0.137	0.115
GRAM	0.538	0.184	0.275	0.237
SVM	**0.797**	0.705	0.748	0.884
BL	0.794	0.748	**0.770**	0.942

all classifiers output word meanings according to the same definition.

Stevenson et al. also proposed a method using a sense-tagged corpus and an MRD for WSD (Stevenson and Wilks, 2001). Although they handled a large vocabulary in their experiment, target words for WSD were restricted to those which appeared in the sense-tagged corpus. Thus the usage of multiple knowledge resources contributed to improving the precision, but not the recall or applicability. Our approach aimed at the improvement of recall and applicability, as indicated in Table 2. Furthermore, the above three methods used different language resources aimed at English WSD, while this paper was concerned with Japanese WSD.

5 Conclusion

In this paper, we proposed a method that combines several classifiers, using different language resources, for Japanese WSD. Two classifiers using an MRD and two classifiers trained on a sense-tagged corpus were combined according to the performance of each classifier on the validation data set. The combined model outperformed the best single classifier for recall and applicability.

In future, we hope to increase the precision of the combined model. The classifiers using example sentences and grammatical information in an MRD should be improved to achieve higher precision. We will conduct an error analysis on these classifiers and investigate ways to improve them.

References

E. Agirre, G. Rigau, L. Padró, and J. Atserias. 2000. Combining supervised and unsupervised lexical knowledge methods for word sense disambiguation. *Computers and the Humanities*, 34(1,2):103–108.

Atsushi Fujii, Kentaro Inui, Takenobu Tokunaga, and Hozumi Tanaka. 1996. To what extent does case contribute to verb sense disambiguation? In *Proceedings of the COLING*, pages 59–64.

Koiti Hasida, Hitoshi Isahara, Takenobu Tokunaga, Minako Hashimoto, Shiho Ogino, Wakako Kashino, Jun Toyoura, and Hironobu Takahashi. 1998. The RWC text databases. In *Proceedings of the LREC*, pages 457–462.

Satoshi Ikehara, Masahiro Miyazaki, Satoshi Shirai, Akio Yokoo, Hiromi Nakaiwa, Kentaro Ogura, Oyama Hiroshi, and Yoshihiko Hayashi. 1997. *Nihongo Goi Taikei (in Japanese)*. Iwanami Shoten, Publishers.

Dan Klein, Kristina Toutanova, H. Tolga Ilhan, Sepandar D. Kamvar, and Christopher D. Manning. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, pages 74–80.

Hand Li and Jun-ichi Takeuchi. 1997. Using evidence that is both strong and reliable in Japanese homograph disambiguation. In *SIG-NL, Information Processing Society of Japan*, pages 53–59.

Kenneth C. Litkowski. 2002. Sense information for disambiguation: Confluence of supervised and unsupervised methods. In *Proceedings of the SIGLEX/SENSEVAL Workshop on Word Sense Disambiguation*, pages 47–53.

Masaki Murata, Masao Utiyama, Kiyotaka Uchimoto, Qing Ma, and Hitoshi Isahara. 2001. Japanese word sense disambiguation using the simple bayes and support vector machine methods. In *Proceedings of the SENSEVAL-2*, pages 135–138.

Minoru Nishio, Etsutaro Iwabuchi, and Shizuo Mizutani. 1994. *Iwanami Kokugo Jiten Dai Go Han*. Iwanami Publisher. (in Japanese).

Ted Pedersen. 2001. A decision tree of bigrams is an accurate predictor of word sense. In *Proceedings of the NAACL*, pages 79–86.

Bernhard Schölkopf. 2000. New support vector algorithms. *Neural Computation*, 12:1083–1121.

Hiroyuki Shinnou and Minoru Sasaki. 2002. Unsupervised learning of word sense disambiguation rules by estimating an optimum iteration number in EM algorithm (in Japanese). In *SIG-NL, Information Processing Society of Japan*, pages 51–58.

Hiroyuki Shinnou. 2002. Learning of word sense disambiguation rules by co-training, checking co-occurrense of features. In *Proceedings of the LREC*, pages 1380–1384.

Mark Stevenson and Yorick Wilks. 2001. The interaction of knowledge sources in word sense disambiguation. *Computational Linguistics*, 27(3):321–349.

Hiroya Takamura et al. 2001. Ensembling based on feature space restructuring with application to WSD. In *Proceedings of NLPRS*, pages 41–48.